3
The Normal Linear Regression Model with Natural Conjugate Prior and Many Explanatory Variables

3.1 INTRODUCTION

In this chapter, we extend the results of the previous chapter to the more reasonable case where the linear regression model has several explanatory variables. The structure of this chapter is very similar to the previous one. The primary difference is that this chapter uses matrix algebra. Despite what many students beginning their study of econometrics might think, matrix algebra is a great simplifier. It offers a useful compact notation for writing out and manipulating formulae and simplifies many derivations. Appendix A offers a very brief introduction to the parts of matrix algebra which will be used in this book. The reader who is unfamiliar with matrix algebra should read this appendix before reading this chapter. Poirier (1995), Greene (2000), or Judge *et al.* (1985) all have good chapters on matrix algebra (and additional references), and the reader is referred to these for further detail.

The steps and derivations in this chapter are, apart from the introduction of matrix algebra, virtually identical to those in the previous chapter. Hence, some readers may find it useful to flip back and forth between this chapter and the previous one. That is, it is easier to understand or motivate derivations or results in matrix form if you first understand them without matrix algebra. Throughout this chapter, we point out similarities between the matrix formulae and their counterparts in the previous chapter as a way of easing the transition to matrix algebra.

3.2 THE LINEAR REGRESSION MODEL IN MATRIX NOTATION

Suppose we have data on a dependent variable, y_i, and k explanatory variables, $x_{i1}, \ldots ,x_{ik}$ for $i = 1, \ldots , N$. The linear regression model is given by

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i \tag{3.1}$$

Our notation is such that x_{i1} is implicitly set to 1 to allow for an intercept. This model can be written more compactly in matrix notation by defining the $N \times 1$ vectors:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_N \end{bmatrix}$$

and

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_N \end{bmatrix}$$

the $k \times 1$ vector

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_k \end{bmatrix}$$

and the $N \times k$ matrix

$$X = \begin{bmatrix} 1 & x_{12} & .. & x_{1k} \\ 1 & x_{22} & .. & x_{2k} \\ . & . & .. & . \\ . & . & .. & . \\ 1 & x_{N2} & .. & x_{Nk} \end{bmatrix}$$

and writing

$$y = X\beta + \varepsilon \tag{3.2}$$

Using the definition of matrix multiplication (see Appendix A, Definition A.4), it can be verified that (3.2) is equivalent to the N equations defined by (3.1).

3.3 THE LIKELIHOOD FUNCTION

The likelihood can be derived in the same manner as in the previous chapter, with the exception that we use matrix notation. Assumptions about ε and X determine the form of the likelihood function. The matrix generalizations of the assumptions in the previous chapter are:

1. ε has a multivariate Normal distribution with mean 0_N and covariance matrix $\sigma^2 I_N$, where 0_N is an N-vector with all elements equal to 0, and I_N is the $N \times N$ identity matrix. Notation for this is: ε is $N(0_N, h^{-1} I_N)$ where $h = \sigma^{-2}$.
2. All elements of X are either fixed (i.e. not random variables) or, if they are random variables, they are independent of all elements of ε with a probability density function, $p(X|\lambda)$, where λ is a vector of parameters that does not include β and h.

The *covariance matrix* of a vector is a matrix that contains the variances of all the elements of the vector on the diagonal and the covariances between different elements filling out the rest of the matrix. In the present context, this means:

$$var(\varepsilon) \equiv \begin{bmatrix} var(\varepsilon_1) & cov(\varepsilon_1, \varepsilon_2) \ldots & cov(\varepsilon_1, \varepsilon_N) \\ cov(\varepsilon_1, \varepsilon_2) & var(\varepsilon_2) \quad \ldots & . \\ . & cov(\varepsilon_2, \varepsilon_3) \ldots & . \\ . & . \qquad \ldots & cov(\varepsilon_{N-1}, \varepsilon_N) \\ cov(\varepsilon_1, \varepsilon_N) & . \qquad \ldots & var(\varepsilon_N) \end{bmatrix}$$

$$= \begin{bmatrix} h^{-1} & 0 & . & . & 0 \\ 0 & h^{-1} & . & . & . \\ . & . & . & . & . \\ . & . & . & . & 0 \\ 0 & . & . & 0 & h^{-1} \end{bmatrix}$$

In other words, the statement that $var(\varepsilon) = h^{-1} I_N$ is a compact notation for $var(\varepsilon_i) = h^{-1}$ and $cov(\varepsilon_i, \varepsilon_j) = 0$ for $i, j = 1, \ldots, N$ and $i \neq j$.

The second assumption implies that we can proceed conditionally on X and treat $p(y|X, \beta, h)$ as the likelihood function. As in the previous chapter, we drop the X from the conditioning set to simplify the notation.

Using the definition of the multivariate Normal density, we can write the likelihood function as:

$$p(y|\beta, h) = \frac{h^{\frac{N}{2}}}{(2\pi)^{\frac{N}{2}}} \left\{ \exp\left[-\frac{h}{2}(y - X\beta)'(y - X\beta) \right] \right\} \tag{3.3}$$

Comparing this equation to (2.2), it can be seen that $(y - X\beta)'(y - X\beta)$ enters in the same manner as $\sum (y_i - \beta x_i)^2$, and it can be confirmed that matrix constructs of the form $a'a$, where a is a vector, are sums of squares.

It proves convenient to write the likelihood function in terms of OLS quantities comparable to (2.3)–(2.5). These are (see Greene (2000), or any other frequentist econometrics textbook which uses matrix algebra):

$$\nu = N - k \tag{3.4}$$

$$\widehat{\beta} = (X'X)^{-1}X'y \tag{3.5}$$

and

$$s^2 = \frac{(y - X\widehat{\beta})'(y - X\widehat{\beta})}{\nu} \tag{3.6}$$

Using a matrix generalization of the derivation in Chapter 2 (see the material between (2.2) and (2.6)), it can be shown that the likelihood function can be written as

$$p(y|\beta, h) = \frac{1}{(2\pi)^{\frac{N}{2}}} \left\{ h^{\frac{1}{2}} \exp\left[-\frac{h}{2}(\beta - \widehat{\beta})' X'X(\beta - \widehat{\beta}) \right] \right\} \left\{ h^{\frac{\nu}{2}} \exp\left[-\frac{h\nu}{2s^{-2}} \right] \right\} \tag{3.7}$$

3.4 THE PRIOR

The form of (3.7) suggests that the natural conjugate prior is Normal-Gamma, and this is indeed the case. In other words, if we elicit a prior for β conditional on h of the form

$$\beta|h \sim N(\underline{\beta}, h^{-1}\underline{V})$$

and a prior for h of the form

$$h \sim G(\underline{s}^{-2}, \underline{\nu})$$

then the posterior will also have these forms. In terms of the notation for the Normal-Gamma distribution, we have

$$\beta, h \sim NG(\underline{\beta}, \underline{V}, \underline{s}^{-2}, \underline{\nu}) \tag{3.8}$$

Note that (3.8) is identical to (2.7), except that $\underline{\beta}$ is now a k-vector containing the prior means for the k regression coefficients, $\beta_1, \ldots, \beta_k$, and $\underline{V}$ is now a $k \times k$ positive definite prior covariance matrix. The notation for the prior density is $p(\beta, h) = f_{NG}(\beta, h|\underline{\beta}, \underline{V}, \underline{s}^{-2}, \underline{\nu})$.

3.5 THE POSTERIOR

The posterior is derived by multiplying the likelihood in (3.7) by the prior in (3.8), and collecting terms (see Exercise 2). Doing so yields a posterior of the form

$$\beta, h|y \sim NG(\overline{\beta}, \overline{V}, \overline{s}^{-2}, \overline{\nu}) \tag{3.9}$$

where

$$\overline{V} = (\underline{V}^{-1} + X'X)^{-1} \tag{3.10}$$

$$\overline{\beta} = \overline{V}(\underline{V}^{-1}\underline{\beta} + X'X\widehat{\beta}) \tag{3.11}$$

$$\overline{\nu} = \underline{\nu} + N \tag{3.12}$$

and $\overline{s}^{-2}$ is defined implicitly through

$$\overline{\nu s}^2 = \underline{\nu s}^2 + \nu s^2 + (\widehat{\beta} - \underline{\beta})'[\underline{V} + (X'X)^{-1}]^{-1}(\widehat{\beta} - \underline{\beta}) \tag{3.13}$$

The previous expressions describe the joint posterior distribution. If we are interested in the marginal posterior for β, we can integrate out h as in Chapter 2 (see (2.13)). The result is a multivariate t distribution. In terms of the notation of Appendix B

$$\beta|y \sim t(\overline{\beta}, \overline{s}^2\overline{V}, \overline{\nu}) \tag{3.14}$$

and it follows from the definition of the t distribution that

$$E(\beta|y) = \overline{\beta} \tag{3.15}$$

and

$$var(\beta|y) = \frac{\overline{\nu s}^2}{\overline{\nu} - 2}\overline{V} \tag{3.16}$$

The properties of the Normal-Gamma distribution imply immediately that:

$$h|y \sim G(\overline{s}^{-2}, \overline{\nu}) \tag{3.17}$$

and, hence, that

$$E(h|y) = \overline{s}^{-2} \tag{3.18}$$

and

$$var(h|y) = \frac{2\overline{s}^{-2}}{\overline{\nu}} \tag{3.19}$$

These expressions are very similar to (2.8)–(2.18), except that now they are written in terms of matrices or vectors instead of scalars. For instance, $\widehat{\beta}$ is now a vector instead of a scalar, the matrix $(X'X)^{-1}$ plays the role that the scalar $\frac{1}{\sum x_i^2}$ did in Chapter 2, $\overline{V}$ is now a $k \times k$ matrix, etc. The interpretation of these formulae is also very similar. For instance, in Chapter 2 we said that the posterior mean of β, $\overline{\beta}$, was a weighted average of the prior mean, $\underline{\beta}$, and the OLS estimate, $\widehat{\beta}$ where the weights reflected the strength of information in the prior ($\underline{V}^{-1}$) and the data ($\sum x_i^2$). Here, the same intuition holds, except the posterior mean is a matrix-weighted average of prior and data information (see also Exercise 6).

The researcher must elicit the prior hyperparameters, $\underline{\beta}, \underline{V}, \underline{s}^{-2}, \underline{\nu}$. In many cases, economic theory, common sense, or a knowledge of previous empirical studies using different data sets will allow her to do so. The fact that the natural conjugate prior can be treated as though it comes from a fictitious data set generated from the same process as the actual data facilitates prior elicitation. Alternatively, the researcher may try a wide range of priors in a prior sensitivity analysis or work with a relatively noninformative prior. For instance, one could set $\underline{\nu}$ to a value much smaller than N and $\underline{V}$ to a 'large' value. When we are dealing with matrices, the interpretation of the term 'large' is not immediately obvious. The matrix generalization of the statement, $a > b$, where a and b are scalars, is usually taken to be $A - B$ is positive definite, where A and B are square matrices. One measure of the magnitude of a matrix is its determinant. Hence, when we say 'A should be large relative to B', we mean that $A - B$ should be a positive definite matrix with large determinant (see Appendix A, Definitions A.10 and A.14 for definitions of the determinant and positive definiteness).

Taking the argument in the previous paragraph to the limit suggests that we can create a purely noninformative prior by setting $\underline{\nu} = 0$ and setting $\underline{V}^{-1}$ to a small value. There is not a unique way of doing the latter (see Exercise 5). One common way is to set $\underline{V}^{-1} = cI_k$, where c is a scalar, and then let c go to zero. If we do this we find $\beta, h|y \sim NG(\overline{\beta}, \overline{V}, \overline{s}^{-2}, \overline{\nu})$, where

$$\overline{V} = (X'X)^{-1} \tag{3.20}$$

$$\overline{\beta} = \widehat{\beta} \tag{3.21}$$

$$\overline{\nu} = N \tag{3.22}$$

and

$$\overline{\nu}\overline{s}^2 = \nu s^2 \tag{3.23}$$

As we found for the simpler model in the previous chapter, all of these formulae involve only data information, and are equal to ordinary least squares quantities.

As for the case with a single explanatory variable, this noninformative prior is improper and can be written as:

$$p(\beta, h) \propto \frac{1}{h} \tag{3.24}$$

3.6 MODEL COMPARISON

The linear regression framework with k explanatory variables allows for a wide variety of models to be compared. In this section, we consider two sorts of model comparison exercise. In the first, models are distinguished according to inequality restrictions on the parameter space. In the second, models are distinguished by equality restrictions.

3.6.1 Model Comparison Involving Inequality Restrictions

In some cases, interest might focus on regions of the parameter space. Consider, for instance, a marketing example where the dependent variable is sales of a product, and one of the explanatory variables reflects spending on a particular advertising campaign. In this case, the econometrician might be interested in finding out whether the advertising campaign increased sales (i.e. whether the coefficient on advertising was positive). In a production example, the econometrician could be interested in finding out whether returns to scale are increasing or decreasing. In terms of a regression model, increasing/decreasing returns to scale could manifest itself as particular combination of coefficients being greater/less than one. Both examples involve an inequality restriction involving one or more of the regression coefficients.

Suppose the inequality restrictions under consideration are of the form:

$$R\beta \geq r \tag{3.25}$$

where R is a known $J \times k$ matrix and r is a known J-vector. Equation (3.25) allows for any J linear inequality restrictions on the regression coefficients, β. To ensure that the restrictions are not redundant, we also must assume $\text{rank}(R) = J$. We can now define two models of the form:

$$M_1 : R\beta \geq r$$

and

$$M_2 : R\beta \ngeq r$$

where the notation in the equation defining M_2 means that one or more of the J inequality restrictions in M_1 are violated.

For models defined in this way, calculating posterior odds ratios is typically quite easy, and the use of noninformative priors is not a problem. That is,

$$PO_{12} = \frac{p(M_1|y)}{p(M_2|y)} = \frac{p(R\beta \geq r|y)}{p(R\beta \ngeq r|y)} \tag{3.26}$$

Since the posterior for β has a multivariate t distribution (see (3.14)), it follows that $p(R\beta|y)$ also has a t distribution (see Appendix B, Theorem B.14). Computer packages such as MATLAB allow for simple calculation of interval probabilities involving the t distribution and, hence, $p(R\beta \geq r|y)$ can easily be calculated. Alternatively, if $J = 1$, statistical tables for the univariate t distribution can be used.

3.6.2 Equality Restrictions

Model comparison involving equality restrictions is slightly more complicated, and additional issues arise with the use of noninformative priors. There are typically two types of model comparison exercise which fall into this category. First, the researcher might be interested in comparing M_1 which imposes $R\beta = r$ to

M_2, which does not have this restriction. M_1 is an example of a model that is *nested* in another model (i.e. M_1 is obtained from M_2 by placing the restrictions $R\beta = r$ on the latter's parameters). Secondly, the researcher might be interested in comparing $M_1 : y = X_1\beta_{(1)}+\varepsilon_1$ to $M_2 : y = X_2\beta_{(2)}+\varepsilon_2$, where X_1 and X_2 are matrices containing completely different explanatory variables. We use the notation $\beta_{(j)}$ to indicate the regression coefficients in the jth model (for $j = 1, 2$), since we have already used the notation β_j to indicate the scalar regression coefficients in (3.1). This is an example of *non-nested* model comparison.[1]

Both these categories involving equality restrictions can be dealt with by writing the two models to be compared as:

$$M_j : y_j = X_j\beta_{(j)} + \varepsilon_j \tag{3.27}$$

where $j = 1, 2$ indicates our two models, y_j will be defined below, X_j is an $N \times k_j$ matrix of explanatory variables, $\beta_{(j)}$ is a k_j-vector of regression coefficients and ε_j is an N-vector of errors distributed as $N(0_N, h_j^{-1} I_N)$.

The case of non-nested model comparison can be dealt with by setting $y_1 = y_2$. The case of nested model comparison involves beginning with the unrestricted linear regression model in (3.2). M_2 is simply this unrestricted model. That is, we set $y_2 = y$, $X_2 = X$ and $\beta_{(2)} = \beta$. M_1, which imposes $R\beta = r$, can be dealt with by imposing the restrictions on the explanatory variables. This may imply a redefinition of the dependent variable. A detailed discussion of how to do this at a high level of generality is given in Poirier (1995, pp. 540–541). However, a consideration of a few examples should be enough to show that restrictions of the form $R\beta = r$ can always be imposed on (3.2) by suitably redefining the explanatory and dependent variables. Restrictions of the form $\beta_m = 0$ imply that X_1 is simply X with the mth explanatory variable omitted. Restrictions of the form $\beta_m = r$ imply that X_1 is simply X with the mth explanatory variable omitted, and $y_1 = y - r x_m$ where x_m is the mth column of X. The restriction $\beta_2 - \beta_3 = 0$ can be handled by deleting the second and third explanatory variables, and inserting a new explanatory variable which is the sum of these deleted variables. Multiple and/or more complicated restrictions can be handled by generalizing the concepts illustrated by these simple examples in a straightforward way.

We denote the Normal-Gamma priors for the two models by:

$$\beta_{(j)}, h_j | M_j \sim NG(\underline{\beta}_j, \underline{V}_j, \underline{s}_j^{-2}, \underline{\nu}_j) \tag{3.28}$$

for $j = 1, 2$. The posteriors take the form

$$\beta_{(j)}, h_j | y_j \sim NG(\overline{\beta}_j, \overline{V}_j, \overline{s}_j^{-2}, \overline{\nu}_j) \tag{3.29}$$

[1] Non-nested model comparison problems can be put into the form of nested model comparison problems by defining M_3, which has explanatory variables $X = [X_1, X_2]$. If we did this, M_1 and M_2 would both be nested in M_3.

where

$$\overline{V}_j = (\underline{V}_j^{-1} + X_j'X_j)^{-1} \tag{3.30}$$

$$\overline{\beta}_j = \overline{V}_j(\underline{V}_j^{-1}\underline{\beta}_j + X_j'X_j\widehat{\beta}_j) \tag{3.31}$$

$$\overline{\nu}_j = \underline{\nu}_j + N \tag{3.32}$$

and $\overline{s}_j^{-2}$ is defined implicitly through

$$\overline{\nu}_j\overline{s}_j^2 = \underline{\nu}_j\underline{s}_j^2 + \nu_j s_j^2 + (\widehat{\beta}_j - \underline{\beta}_j)'[\underline{V}_j + (X_j'X_j)^{-1}]^{-1}(\widehat{\beta}_j - \underline{\beta}_j) \tag{3.33}$$

$\widehat{\beta}_j$, s_j^2 and ν_j are OLS quantities analogous to (3.4)–(3.6).

The derivation of the marginal likelihood for each model and, hence, the posterior odds ratio, proceeds along the same lines as in the previous chapter (see (2.31)–(2.34)). In particular, the marginal likelihood becomes

$$p(y_j|M_j) = c_j\left(\frac{|\overline{V}_j|}{|\underline{V}_j|}\right)^{\frac{1}{2}}(\overline{\nu}_j\overline{s}_j^2)^{-\frac{\overline{\nu}_j}{2}} \tag{3.34}$$

for $j = 1, 2$, where

$$c_j = \frac{\Gamma\left(\frac{\overline{\nu}_j}{2}\right)(\underline{\nu}_j\underline{s}_j^2)^{\frac{\underline{\nu}_j}{2}}}{\Gamma\left(\frac{\underline{\nu}_j}{2}\right)\pi^{\frac{N}{2}}} \tag{3.35}$$

The posterior odds ratio comparing M_1 to M_2 is

$$PO_{12} = \frac{c_1\left(\frac{|\overline{V}_1|}{|\underline{V}_1|}\right)^{\frac{1}{2}}(\overline{\nu}_1\overline{s}_1^2)^{-\frac{\overline{\nu}_1}{2}}p(M_1)}{c_2\left(\frac{|\overline{V}_2|}{|\underline{V}_2|}\right)^{\frac{1}{2}}(\overline{\nu}_2\overline{s}_2^2)^{-\frac{\overline{\nu}_2}{2}}p(M_2)} \tag{3.36}$$

The factors which affect the posterior odds ratio were discussed in Chapter 2. In particular, the posterior odds ratio depends upon the prior odds ratio, and contains rewards for model fit, coherency between prior and data information and parsimony.

The issue of the reward for parsimony relates closely to problems involved with use of noninformative priors. When discussing posterior inference, we considered a prior where $\underline{\nu} = 0$ and $\underline{V}^{-1} = cI_k$, where c was a scalar. We then defined a noninformative prior as one where c was set to zero. Loosely speaking, setting $\underline{\nu} = 0$ implies there is no prior information about the error precision, h, and letting c go to zero implies there is no prior information about the regression coefficients, β. In this section, we consider these two steps for becoming noninformative separately. An important result will be that it is reasonable to use noninformative priors for h_j for $j = 1, 2$, but it is not reasonable to use noninformative priors for $\beta_{(j)}$. The reason is that the error precision is a parameter which is common to both models, and has the same interpretation in each. However, $\beta_{(1)}$ and $\beta_{(2)}$

are not the same and, in cases where $k_1 \neq k_2$, the use of noninformative priors causes serious problems for Bayesian model comparison using a posterior odds ratio. These considerations motivate an important rule of thumb: *When comparing models using posterior odds ratios, it is acceptable to use noninformative priors over parameters which are common to all models. However, informative, proper priors should be used over all other parameters*. This rule of thumb is relevant not only for the regression model, but for virtually any model you might wish to use.

To justify the statements in the previous paragraph, consider first what happens if we set $\underline{\nu}_1 = \underline{\nu}_2 = 0$.[2] The formula for the posterior odds ratio in (3.36) simplifies substantially since $c_1 = c_2$. However, the posterior odds ratio still has a sensible interpretation involving model fit (i.e. s_j^2), the coherency between prior and data information (see the last term in (3.33)), etc. In short, using a noninformative prior for the error precisions in the two models is perfectly reasonable.

However, using noninformative priors for the $\beta_{(j)}$'s causes major problems which occur largely when $k_1 \neq k_2$. In the case of non-nested model comparison, we have two models which have different explanatory variables and it is clear that the dimension and interpretation of $\beta_{(1)}$ and $\beta_{(2)}$ can be different. For the case of nested model comparison, the restrictions imposed under M_1 will ensure that $\beta_{(1)}$ is of lower dimension than $\beta_{(2)}$ and, hence, $k_1 < k_2$. Thus, having $k_1 \neq k_2$ is quite common. The problem with interpreting posterior odds ratios in this case occurs because of the term $|\underline{V}_j|$. If we set $\underline{V}_j^{-1} = cI_{k_j}$, then $|\underline{V}_j| = \frac{1}{c^{k_j}}$. If we then let c go to zero, an examination of (3.36) should convince you that terms involving c will not cancel out. In fact, provided the prior odds ratio is positive and finite, if $k_1 < k_2$, PO_{12} becomes infinite, while if $k_1 > k_2$, PO_{12} goes to zero. In other words, the posterior odds ratio will always lend overwhelming support for the model with fewer parameters, regardless of the data. In the limit, the reward for parsimony becomes completely dominant and the more parsimonious model is always selected! Clearly, this is unreasonable and provides a strong argument for saying that informative priors should always be used for $\beta_{(1)}$ and $\beta_{(2)}$, at least for coefficients that are not common to both models.

You may think that you are safe when $k_1 = k_2$ as, in this case, the noninformative prior yields a posterior odds ratio of:

$$PO_{12} = \frac{(|X_1'X_1|)^{\frac{1}{2}}(\nu_1 s_1^2)^{-\frac{N}{2}}p(M_1)}{(|X_2'X_2|)^{\frac{1}{2}}(\nu_2 s_2^2)^{-\frac{N}{2}}p(M_2)} \tag{3.37}$$

Note, however, that this expression depends upon units of measurement. For instance, if your explanatory variables in M_1 are measured in dollars and you decide to change this and measure them in thousands of dollars, leaving X_2 unchanged, your posterior odds ratio will change. This is a very undesirable feature which makes many Bayesians reluctant to use posterior odds based on

[2] To be mathematically precise, we should let them go to zero at the same rate.

noninformative priors, even in the case where $k_1 = k_2$. When the researcher elicits an informative prior, this problem does not arise. For instance, in the empirical illustration in the next section, the dependent variable is house price (measured in dollars) and one of the explanatory variables, x_2, is the lot size (measured in square feet). The coefficient, β_2, can be interpreted through a statement of the form: "An extra square foot of lot size will tend to add β_2 dollars to the price of a house, holding other house characteristics constant". The researcher would choose a prior for β_2 with this interpretation in mind. However, if the units of measurement of x_2 were changed to hundreds of square feet, then the interpretation would be based on a statement of the form: "An extra hundred square feet of lot size will tend to add β_2 dollars to the price of a house, holding other house characteristics constant". β_2 has a very different interpretation if the units of measurement of x_2 are changed and, hence, the researcher would use a very different prior. In other words, when the researcher elicits an informative prior, she is implicitly taking into account the units of measurement. However, with a noninformative prior the researcher does not take into account such considerations.

An important message of this section is that, when doing model comparison, it is important to elicit informative priors for parameters which differ or are restricted across models. With the other activities that an econometrician might do (i.e. estimation and prediction) noninformative priors are an acceptable path to take for the Bayesian who seeks to remain 'objective' and not introduce prior information. However, when calculating posterior odds ratios, a noninformative path may not be acceptable.

The ideas in this section have all been developed for the case of two models but can be extended to the case of many models in a straightforward way (see the discussion after (1.7) in Chapter 1). We also stress that posterior odds ratios can be used to form the posterior model probabilities which are necessary for Bayesian model averaging (see (2.42)).

3.6.3 Highest Posterior Density Intervals

Standard Bayesian model comparison techniques are based on the intuitively appealing idea that $p(M_j|y)$ summarizes all of our knowledge and uncertainty about M_j after seeing the data. However, as we have seen, calculating meaningful posterior model probabilities typically requires the elicitation of informative priors. For the Bayesian who wants to do model testing or comparison with a noninformative prior, there are some other techniques which can be used. However, these techniques are not as intuitively appealing as Bayesian model probabilities and have only *ad hoc* justifications. In later chapters, we discuss some of these techniques. In this subsection, we introduce the idea of a Highest Posterior Density Interval (HPDI), and show how it can be used in an *ad hoc* fashion to compare nested models.

Before discussing model comparison, we begin with some definitions of basic concepts. We define these concepts in the context of the parameter vector β in the Normal linear regression model, but they are quite general and can be used with the parameters of any model. Suppose that the elements of the vector of regression coefficients, β, can each lie anywhere in the interval $(-\infty, \infty)$, which is denoted by $\beta \in R^k$. Let $\omega = g(\beta)$ be some m-vector of functions of β which is defined over a region, Ω, where $m \leq k$. Let C be a region within Ω, denoted by $C \subseteq \Omega$.

Definition 3.1: Credible Sets

The set $C \subseteq \Omega$ is a $100(1-\alpha)\%$ credible set with respect to $p(\omega|y)$ if:

$$p(\omega \in C|y) = \int_C p(\omega|y)d\omega = 1 - \alpha$$

As an example, suppose $\omega = g(\beta) = \beta_j$, a single regression coefficient. Then a 95% credible interval for β_j is any interval, $[a, b]$ such that:

$$p(a \leq \beta_j \leq b|y) = \int_a^b p(\beta_j|y)d\beta_j = 0.95$$

There are typically numerous possible credible intervals. Suppose, for instance, that $\beta_j|y$ is $N(0, 1)$. Then, using statistical tables for the standard Normal, we find that $[-1.96, 1.96]$ is a 95% credible interval, as is $[-1.75, 2.33]$ and $[-1.64, \infty)$, etc. To choose from among the infinite number of credible intervals, it is common to choose the one with smallest area. In the standard Normal example, $[-1.96, 1.96]$ is the shortest credible interval. The name given for such a choice is a *Highest Posterior Density Interval*. This is formalized in the following definition.

Definition 3.2: Highest Posterior Density Intervals

A $100(1-\alpha)\%$ highest posterior density interval for ω is a $100(1-\alpha)\%$ credible interval for ω with the property that it has a smaller area than any other $100(1-\alpha)\%$ credible interval for ω.

It is common to present highest posterior density intervals in addition to point estimates when doing Bayesian estimation. For instance, the researcher might report a posterior mean plus a 95% HPDI of β_j. The researcher is 95% sure that β_j lies within the HPDI. HPDIs can also be used in an *ad hoc* manner to do model comparison. Consider, for instance, two Normal linear regression models as in (3.2), and that interest centers on deciding whether the jth explanatory variable should be included. Thus, the two models under consideration are

$$M_1 : \beta_j = 0$$

and

$$M_2 : \beta_j \neq 0$$

Posterior inference under M_2 can be performed as outlined in (3.28)–(3.33), and an HPDI can be calculated for β_j using the properties of the t distribution. If this HPDI does not include zero, then this is evidence against M_1. A finding that the HPDI does include zero is taken as evidence in favor of M_1. Such a strategy can be generalized in the obvious way to the case where we are interested in investigating whether $R\beta = r$.

The reader who knows frequentist econometrics will recognize the similarity of this approach with common hypothesis testing procedures. For instance, a frequentist test of the hypothesis that $\beta_j = 0$ can be done by calculating a confidence interval for β_j. If this confidence interval contains zero, then the hypothesis is accepted. If it does not, the hypothesis is rejected. We stress, however, that this similarity only holds far enough to provide some very crude intuition. Confidence intervals have a very different interpretation from HPDIs.

HPDIs are a very general tool in that they will exist any time the posterior exists. Thus, they can be used with the noninformative prior discussed previously. However, the justification for using them to compare models, although sensible, is an informal one which, in contrast to posterior odds, is not rooted firmly in probability theory.

3.7 PREDICTION

Prediction for the case of the Normal linear regression model with a single explanatory variable is outlined in Chapter 2 (2.36)–(2.40). The case of several explanatory variables is a simple extension of this material. Suppose we have a Normal linear regression model as in (3.2), with likelihood and prior given in (3.3) and (3.8). Posterior inference can be carried out using (3.9). We want to carry out predictive inference on T unobserved values of the dependent variable, which we denote by $y^* = (y_1^*, \ldots, y_T^*)'$, which are generated according to

$$y^* = X^*\beta + \varepsilon^* \tag{3.38}$$

where ε^* is independent of ε and is $N(0, h^{-1}I_T)$ and X^* is a $T \times k$ matrix analogous to X, containing the k explanatory variables for each of the T out-of-sample data points.

The steps in deriving the predictive density for y^* are simple generalizations of those outlined in (2.37)–(2.40). That is, for the Bayesian prediction is based on

$$p(y^*|y) = \int\int p(y^*|y, \beta, h)p(\beta, h|y)d\beta\, dh$$

The fact that ε^* is independent of ε implies that y and y^* are independent of one another and, hence, $p(y^*|y, \beta, h) = p(y^*|\beta, h)$. The latter term can be written as

$$p(y^*|\beta, h) = \frac{h^{\frac{S}{2}}}{(2\pi)^{\frac{S}{2}}} \exp\left[-\frac{h}{2}(y^* - X^*\beta)'(y^* - X^*\beta)\right] \tag{3.39}$$

Multiplying (3.38) by the posterior given in (3.9) and integrating yields a multivariate t predictive density of the form

$$y^*|y \sim t(X^*\overline{\beta}, \overline{s}^2\{I_T + X^*\overline{V}X^{*'}\}, \overline{\nu}) \tag{3.40}$$

This result can be used to carry out predictive inference in the Normal linear regression model with natural conjugate prior.

3.8 COMPUTATIONAL METHODS: MONTE CARLO INTEGRATION

Model comparison, prediction and posterior inference about β can all be done analytically using the results in previous sections. Furthermore, since the marginal posterior for β is a multivariate t distribution, linear combinations of β are also multivariate t (see Appendix B, Theorem B.14). Thus, if R is defined as in (3.25), posterior inference on $R\beta$ can be carried out using the multivariate t distribution. Since the marginal posterior of h is Gamma, the properties of this well-known distribution can be used to make inferences about the error precision.

However, there are some cases where interest centers not on β, nor on $R\beta$, but on some nonlinear function of β which we will call $f(\beta)$. We will assume $f(.)$ is a scalar function, but the techniques in this section can be extended to several functions by simply handling one function at a time.

In general, the posterior for $f(\beta)$ will not lie in the class of densities with well-known analytical properties. This, then, is a convenient place to start discussing posterior simulation. As described in Chapter 1, even if we do not know the properties (e.g. mean, standard deviation, etc.) of a density, it is possible to figure them out on the computer using simulation. The simplest algorithm for doing posterior simulation is called Monte Carlo integration. In the context of the Normal linear regression model, we can write the basic theorem underlying Monte Carlo integration (see Theorem 1.1) as:

Theorem 3.1: Monte Carlo Integration

Let $\beta^{(s)}$ for $s = 1, \ldots, S$ be a random sample from $p(\beta|y)$ and $g(.)$ be any function and define

$$\widehat{g}_S = \frac{1}{S}\sum_{r=1}^{S} g(\beta^{(s)}) \tag{3.41}$$

then $\widehat{g}_S$ converges to $E[g(\beta)|y]$ as S goes to infinity.

Do not be confused by the introduction of two functions $f(.)$ and $g(.)$. By setting $g(.) = f(.)$, we can obtain an estimate of $E[f(\beta)|y]$ for any $f(.)$. However, we may wish to calculate other posterior properties of $f(\beta)$ and this requires the introduction of the function $g(.)$. For instance, the calculation of $var[f(\beta)|y]$ involves setting $g(.) = f(.)^2$ and using (3.41) to calculate $E[f(\beta)^2|y]$. As described in Chapter 1, by suitably redefining $g(.)$ we can calculate a variety of posterior properties of our function of interest, $f(.)$.

Equation (3.41) says that, given random draws from the posterior for β, inference about any function of the parameters can be done. Here Monte Carlo integration requires computer code which takes random draws from the multivariate t distribution. This is available in many places. For instance, MATLAB code relating to the following empirical illustration is available on the website associated with this book. This shows how Monte Carlo integration is done in practice. The structure of the code is as follows:

Step 1: Take a random draw, $\beta^{(s)}$ from the posterior for β given in (3.14) using a random number generator for the multivariate t distribution.
Step 2: Calculate $g(\beta^{(s)})$ and keep this result.
Step 3: Repeat Steps 1 and 2 S times.
Step 4: Take the average of the S draws $g(\beta^{(1)}), \ldots, g(\beta^{(S)})$.

These steps will yield an estimate of $E[g(\beta)|y]$ for any function of interest.

It is worth stressing that Monte Carlo integration yields only an approximation for $E[g(\beta)|y]$ (since you cannot set $S = \infty$). However, by selecting S, the researcher can control the degree of approximation error. Furthermore, as described in Chapter 1 (see (1.13)), we can obtain a numerical measure of the approximation error using a central limit theorem. In particular, we obtain

$$\sqrt{S}\{\widehat{g}_S - E[g(\beta)|y]\} \to N(0, \sigma_g^2) \tag{3.42}$$

as S goes to infinity, where $\sigma_g^2 = var(g(\beta)|y)$. The latter quantity can itself be estimated using Monte Carlo integration, and we shall call such an estimate $\widehat{\sigma}_g^2$. Using this estimate, (3.42) and the properties of the Normal density we can write:

$$Pr\left\{E[g(\beta)|y] - 1.96\frac{\widehat{\sigma}_g}{\sqrt{S}} \leq \widehat{g}_S \leq E[g(\beta)|y] + 1.96\frac{\widehat{\sigma}_g}{\sqrt{S}}\right\} \approx 0.95 \tag{3.43}$$

We can then rearrange the probability statement in (3.43) to find an approximate 95% confidence interval for $E[g(\beta)|y]$ of the form $\left[\widehat{g}_S - 1.96\frac{\widehat{\sigma}_g}{\sqrt{S}}, \widehat{g}_S + 1.96\frac{\widehat{\sigma}_g}{\sqrt{S}}\right]$. The researcher can present this as a measure of how accurate her estimate of $E[g(\beta)|y]$ is or to use it as a guide for selecting S. Alternatively, the numerical standard error, $\frac{\widehat{\sigma}_g}{\sqrt{S}}$, can be reported as implicitly containing the same information in a more compact form.

3.9 EMPIRICAL ILLUSTRATION

To illustrate Bayesian inference in the multiple regression model, we use a data set containing the sales price of $N = 546$ houses sold in Windsor, Canada in 1987. Further details about this data set are provided in Anglin and Gencay

(1996). Interest centers on finding out which factors affect house prices and, hence, sales price is our dependent variable. We use four explanatory variables: the size of the lot the house is on, the number of bedrooms, number of bathrooms and number of storeys. Thus, we have:

- y_i = sales price of the ith house measured in Canadian dollars,
- x_{i2} = the lot size of the ith house measured in square feet,
- x_{i3} = the number of bedrooms in the ith house,
- x_{i4} = the number of bathrooms in the ith house,
- x_{i5} = the number of storeys in the ith house.

Presumably, a researcher doing work with this data set would have knowledge of the Windsor real estate market, and could use such knowledge to elicit a reasonable informative prior. Or, the researcher could ask a local real estate agent to help provide prior information. For instance, the researcher could ask the real estate agent a series of questions of the form: "How much would you expect a house with a lot of size 4000 square feet, with two bedrooms, one bathroom and one storey to cost?"; "How much would you expect a house with a lot of size 6000 square feet, with three bedrooms, two bathrooms and two storeys to cost?", etc. Since there are five unknown regression coefficients, the answers to five questions of this form would give the researcher five equations in five unknowns. She could then solve these equations to find the real estate agent's implicit guesses as to what the regression coefficients are. These guesses could be used as the prior mean for β.

For illustrative purposes, here we will use only a crudely elicited informative prior. House prices in Windsor in 1987 showed a wide variation, but most houses sold for prices in the \$50 000–\$150 000 region. A regression model which fits well might have errors that typically are of the order of magnitude of a few thousand dollars and maybe \$10 000 at most. This suggests that σ might be roughly 5000. That is, since the errors are Normally distributed with mean zero, if $\sigma = 5000$ then 95% of the errors will be less than $1.96 \times 5000 = \$9800$ in absolute value. Since $h = \frac{1}{\sigma^2}$, this suggests that a reasonable prior guess for h would be $\frac{1}{5000^2} = 4.0 \times 10^{-8}$. Thus, we set $\underline{s}^{-2} = 4.0 \times 10^{-8}$. However, this is a very crude guess and, hence, we want to attach little weight to it by setting $\underline{\nu}$ to a value which is much smaller than N. Since $N = 546$, setting $\underline{\nu} = 5$ is relatively noninformative. Loosely speaking, we are saying our prior information about h should have about 1% of the weight as the data information $\left(\text{i.e. } \frac{\underline{\nu}}{N} \approx 0.01\right)$.

For the regression coefficients, we set:

$$\underline{\beta} = \begin{bmatrix} 0.0 \\ 10 \\ 5000 \\ 10\,000 \\ 10\,000 \end{bmatrix}$$

Remember that regression coefficients can be interpreted as saying "if explanatory variable j is increased by one unit and all other explanatory variables are held constant, the price of the house tends to increase by β_j dollars". Hence, our prior mean implies statements of the form "if we compare two houses which are identical except the first house has one bedroom more than the second, then we expect the first house to be worth \$5000 more than the second" or "if the number of bathrooms is increases by one, holding all other house characteristics constant, we expect the price of the house go up by \$10 000", etc.

All these guesses about the regression coefficients are rather crude, so it makes sense to attach a relatively large prior variance to each of them. For instance, suppose our prior information about the intercept is very uncertain. In this case, we might want $var(\beta_1) = 10\,000^2$ (i.e. the prior standard deviation is 10 000 and, hence, we are attaching approximately 95% prior probability to the region $[-20\,000, 20\,000]$ which is a very wide interval).[3] If we think it highly probable that the effect of lot size would be between 0 and 20, we would choose $var(\beta_2) = 25$ (i.e. choose a prior standard deviation for β_2 of 5). For the other regression coefficients, we choose $var(\beta_3) = 2500^2$ and $var(\beta_4) = var(\beta_5) = 5000^2$. These hyperparameter values say, for instance, that our best prior guess of β_4 is 10 000 and we think it very likely that it lies in the interval $[0, 20\,000]$.

Given these choices, we can figure out the prior covariance matrix. The properties of the Normal-Gamma distribution imply that the prior covariance matrix for β has the form:

$$var(\beta) = \frac{\underline{\nu}\underline{s}^2}{\underline{\nu} - 2}\underline{V}$$

Since $\frac{\underline{\nu}\underline{s}^2}{\underline{\nu}-2} = 41\,666\,666\frac{2}{3}$, our choices for $var(\beta_j)$ for $j = 1, \ldots, 5$ imply:

$$\underline{V} = \begin{bmatrix} 2.40 & 0 & 0 & 0 & 0 \\ 0 & 6.0 \times 10^{-7} & 0 & 0 & 0 \\ 0 & 0 & 0.15 & 0 & 0 \\ 0 & 0 & 0 & 0.60 & 0 \\ 0 & 0 & 0 & 0 & 0.60 \end{bmatrix}$$

Note that we have set all the prior covariances to zero. This is commonly done, since it is often hard to make reasonable guesses about what they might be. It implies that your prior information about what plausible values for β_j might be are uncorrelated with those for β_i for $i \neq j$. In many cases, this is a reasonable assumption. This completes our specification of an informative natural conjugate prior for the parameters of our model.

The preceding paragraphs illustrate how prior elicitation might be done in practice. As you can see, prior elicitation can be a bit complicated and involve a lot of

[3] Here we are using a useful approximate rule-of-thumb that says that roughly 95% of the probability in a density is located within two standard deviations of its mean. This approximation works best for the Normal distribution or distributions which have a similar shape to the Normal (e.g. the t distribution).

guesswork. However, it is a very useful exercise to carry out, since it forces the researcher to think carefully about her model and how its parameters are interpreted. For the researcher who has no prior information (or does not wish to use it), is also possible to carry out a noninformative Bayesian analysis using (3.24).

Tables 3.1 and 3.2 present prior and posterior results for both the informative and noninformative priors. Posterior results based on the informative prior can be calculated using (3.9)–(3.19), and those based on the noninformative prior use (3.20)–(3.23). Table 3.1 confirms that our prior is relatively noninformative, since posterior results based on the informative prior are quite similar to those based on the noninformative prior. In the previous chapter, we saw that the posterior mean of the single regression coefficient using the informative prior lay between the prior mean and the OLS estimate. In Table 3.1, there is also a tendency for the posterior mean based on the informative prior to lie between the prior mean and the OLS estimate. Remember that the OLS estimate is identical to the posterior mean based on the noninformative prior (see (3.21)). However, not every posterior mean based on the informative prior lies between the prior mean and the OLS estimate (see results for β_1). This is because the posterior mean is a matrix weighted average of the prior mean and the OLS estimate (see (3.11)). The matrix weighting does not imply that every individual coefficient lies between its prior mean and OLS estimate.

Table 3.2 presents prior and posterior results for h. For this parameter, too, it can be seen that data information dominates prior information. That is, posterior results using the informative prior are quite similar to those using the noninformative prior.

A written summary of results in Tables 3.1 and 3.2 proceeds in the standard way, based on the interpretation of regression parameters. For instance, the researcher might write statements such as: “Regardless of whether we use the informative or noninformative priors, we find the posterior mean of β_4 to be roughly 17 000. Thus, our point estimate indicates that, if we compare two houses which are the same except the first house has one more bathroom than the second, we would expect the first house to be worth roughly \$17 000 more than the second.” Or, more tersely, “the point estimate of the the marginal effect of bathrooms on house price is roughly \$17 000”.

Table 3.3 contains results relating to the various methods of model comparison discussed in this chapter. All results can be used to shed light on the question of whether an individual regression coefficient is equal to zero. The column labelled $p(\beta_j > 0|y)$ uses (3.14) and the properties of the t distribution to calculate the probability that each individual coefficient is positive. The usefulness of such probabilities is described in Section 3.6.1. The column labelled ‘Posterior Odds in Favor of $\beta_j = 0$’ contains the posterior odds ratio comparing a model which restricts the appropriate element of β to be zero against the unrestricted alternative. That is, it uses the methods outlined in Section 3.6.2 to calculate the posterior odds ratio comparing two regression models:

Table 3.1 Prior and Posterior Means for β (standard deviations in parentheses)

	Prior	Posterior	
	Informative	Using Noninformative Prior	Using Informative Prior
β_1	0	−4009.55	−4035.05
	(10 000)	(3593.16)	(3530.16)
β_2	10	5.43	5.43
	(5)	(0.37)	(0.37)
β_3	5000	2824.61	2886.81
	(2500)	(1211.45)	(1184.93)
β_4	10 000	17 105.17	16 965.24
	(5000)	(1729.65)	(1708.02)
β_5	10 000	7634.90	7641.23
	(5000)	(1005.19)	(997.02)

Table 3.2 Prior and Posterior Properties of h

	Prior	Posterior	
	Informative	Using Noninformative Prior	Using Informative Prior
Mean	4.0×10^{-8}	3.03×10^{-9}	3.05×10^{-9}
St. Deviation	1.6×10^{-8}	3.33×10^{-6}	3.33×10^{-6}

$M_1 : \beta_j = 0$ to $M_2 : \beta_j \neq 0$. The restricted model uses an informative prior which is identical to the unrestricted prior, except that $\underline{\beta}$ and $\underline{V}$ become 4×1 and 4×4 matrices, respectively, with prior information relating to β_j omitted. A prior odds ratio of one is used. The last two columns of Table 3.3. present 99% and 95% Highest Posterior Density Intervals for each β_j using the noninformative prior. As described in Section 3.6.3, HPDIs can be used to carry out tests of equality restrictions. Remember that these have a sensible, but *ad hoc*, justification even when a noninformative prior is used. Don't forget that posterior odds ratios usually require the use of informative priors (at least over parameters which are common to the two models being compared). Hence, we do not present posterior odds ratios using the noninformative prior.

Table 3.3 Model Comparison Involving β

	Informative Prior			
	$p(\beta_j > 0\|y)$	95% HPDI	99% HPDI	Posterior Odds for $\beta_j = 0$
β_1	0.13	[−10 957, 2887]	[−13 143, 5073]	4.14
β_2	1.00	[4.71, 6.15]	[4.49, 6.38]	2.25×10^{-39}
β_3	0.99	[563.5, 5210.1]	[−170.4, 5944]	0.39
β_4	1.00	[13 616, 20 314]	[12 558, 21 372]	1.72×10^{-19}
β_5	1.00	[5686, 9596]	[5069, 10 214]	1.22×10^{-11}
	Noninformative Prior			
	$p(\beta_j > 0\|y)$	95% HPDI	99% HPDI	Posterior Odds for $\beta_j = 0$
β_1	0.13	[−11 055, 3036]	[−13 280, 5261]	—
β_2	1.00	[4.71, 6.15]	[4.48, 6.38]	—
β_3	0.99	[449.3, 5200]	[−301.1, 5950]	—
β_4	1.00	[13 714, 20 497]	[12 642, 21 568]	—
β_5	1.00	[5664, 9606]	[5041, 10 228]	—

The results in Table 3.3 are consistent with those in Table 3.1. In the latter table, we saw that the posterior means of β_2, β_4 and β_5 were all positive and very large relative to their posterior standard deviations, providing strong evidence that all these coefficients are non-zero and positive. Regardless of whether we use the informative or noninformative priors, Table 3.3 indicates $p(\beta_j > 0|y)$ is 1 (to several decimal places) for $j = 2, 4, 5$, and none of the HPDIs contains 0. For the informative prior, the posterior odds ratios comparing $M_1 : \beta_j = 0$ to $M_2 : \beta_j \neq 0$ for $j = 2, 4, 5$, are all very small, indicating that the unrestricted model receives massively more probability than the restricted model. Results for β_1 and β_3 are more mixed. For instance, most of the evidence indicates that $\beta_3 \neq 0$. However, the 99% HPDI for this parameter does include zero. Hence, if we were to use the model selection strategy outlined in Section 3.6.3, our results would depend upon precisely which HPDI we chose. A 95% HPDI would imply that $\beta_3 \neq 0$, whereas the 99% HPDI would imply $\beta_3 = 0$. This uncertainty is reflected in the posterior odds ratio, which indicates that the restricted model is 0.39 times as likely as the unrestricted model. If we use this posterior odds ratio to calculate a posterior model probability we find that $p(M_1 : \beta_3 = 0|y) = 0.28$. In words, there is a 28% chance that $\beta_3 = 0$ and 72% chance that it is not. When such uncertainty is present, it may make sense to consider Bayesian model averaging. The alternative is to choose either the unrestricted or the restricted model. In either case, there is a substantial probability that you are choosing the wrong model.

To illustrate how prediction can be done using the Normal linear regression model, we consider the case where the researcher is interested in predicting the sales price of a house with a lot size of 5000 square feet, two bedrooms, two bathrooms and one storey. Using (3.40), we can work out that the predictive distribution in the case of the informative prior is $t(70\,468, 3.33 \times 10^8, 551)$. For the noninformative prior, the predictive distribution is $t(70\,631, 3.35 \times 10^8, 546)$. The researcher might use either of these predictive densities to present information to a client wishing to sell a house with the characteristics listed above. For instance, she might say that her best guess of the sales price is slightly over $70 000, but that there is a large uncertainty associated with this guess (i.e. the predictive standard deviation is roughly $18 000).

Section 3.8 introduces Monte Carlo integration. As discussed in that section, Monte Carlo integration is not required for the Normal linear regression model with natural conjugate prior, unless interest centers on nonlinear functions of the regression coefficients. That is, we already know the posterior properties of β (see Table 3.1), so there is no need to do Monte Carlo integration here. However, to illustrate how Monte Carlo integration is carried out, we will use it to calculate the posterior mean and standard deviation of β_2. From Table 3.1, we know that these should be 5.43 and 0.37, respectively. This gives us a benchmark to see how well Monte Carlo integration works. For the sake of brevity, we calculate results only for the informative prior.

Monte Carlo integration can be implemented by taking random draws from the posterior distribution of β and then averaging appropriate functions of these draws (see (3.41)). From (3.14), we know that the $p(\beta|y)$ is a t density. Thus, we can write a program which repeatedly takes random draws from (3.14) and averages them.

Table 3.4 presents the posterior mean and standard deviation for β_2 calculated in various ways. The row labelled 'Analytical' is the exact result obtained using (3.14)–(3.16). The other rows present results calculated using Monte Carlo integration with different numbers of replications. These rows also present numerical standard errors (see the discussion at end of Section 3.8) which give insight into the accuracy of the Monte Carlo approximation of $E(\beta_2|y)$.

As expected, the accuracy of approximation of both the posterior mean and standard deviation gets better and better as the number of replications is increased.[4] In an empirical context, the exact choice of S will depend upon the accuracy desired by the researcher. For instance, if the researcher is doing a preliminary exploration of the data, then perhaps a rough estimate will do and setting $S = 10$ or 100 may be enough. However, to get highly accurate estimates (perhaps for the final results written up in a report), then the researcher may set $S = 10\,000$ or even 100 000. The numerical standard error does seem to give a good idea

[4] We remind the reader that the computer programs for calculating the results in the empirical illustrations are available on the website associated with this book. If you use these programs (or create your own programs), you should be able to exactly reproduce all tables up to and including Table 3.3. However, since Monte Carlo integration involves taking random draws, you will not be able to exactly reproduce Table 3.4. That is, your random draws will be different from mine and, hence, your results may differ slightly from mine. Formally, the random generator requires what is called a *seed* to get started. The seed is a number and it is usually taken from the computer's clock. Hence, programs run at different times will yield different random draws.

Table 3.4 Posterior Results for β_2 Calculated Various Ways

	Mean	Standard Deviation	Numerical Standard Error
Analytical	5.4316	0.3662	—
Number of Replications			
$S = 10$	5.3234	0.2889	0.0913
$S = 100$	5.4877	0.4011	0.0401
$S = 1000$	5.4209	0.3727	0.0118
$S = 10\,000$	5.4330	0.3677	0.0037
$S = 100\,000$	5.4323	0.3664	0.0012

of the accuracy of each approximation in that approximate posterior means are rarely much more than one numerical standard error from the true posterior mean given in the row labelled ‘Analytical’.

It is also worth noting that, although increasing S will increase the accuracy of the Monte Carlo approximation of $E(\beta_2|y)$, the increase is not linear in S. For instance, Table 3.4 shows that results with $S = 100\,000$ are not ten times as accurate as those with $S = 10\,000$. Analytically, the numerical standard error, $\frac{\widehat{\sigma}_g}{\sqrt{S}}$, decreases at a rate of $\frac{1}{\sqrt{S}}$. Thus, results with $S = 100\,000$ should only be roughly $\sqrt{10} = 3.16$ times as accurate as those with $S = 10\,000$.

3.10 SUMMARY

In this chapter, we have gone through a complete Bayesian analysis (i.e. likelihood, prior, posterior, model comparison and prediction) for the Normal linear regression model with natural conjugate prior and k explanatory variables. This chapter is mostly the same as the previous one, except that matrix notation is used throughout to accommodate the complications caused by $k > 1$ explanatory variables. The concept of a highest posterior density interval was introduced. We also showed how Monte Carlo integration, a topic first discussed in Chapter 1, can be used to carry out posterior inference on nonlinear functions of the regression parameters.

3.11 EXERCISES

3.11.1 Theoretical Exercises

1. For the Normal linear regression model, show that the likelihood function in (3.3) can be written in terms of OLS quantities as in (3.7).